Global superstore customer analysis

Lora YANG

2022-04-05

About Dataset

Shopping online is currently the need of the hour. Because of this COVID, it's not easy to walk in a store randomly and buy anything you want. I this I am trying to understand a few things like

Business task

Customers Analysis

–Profile the customers based on their frequency of purchase - calculate frequency of purchase for each customer

–Do the high frequent customers are contributing more revenue

–Are they also profitable - what is the profit margin across the buckets

–Which customer segment is most profitable in each year.

–How the customers are distributed across the countries

prepare data

data download from Kaggle

Global_superstore2.csv

Sort, filter and clean the Data use spreedsheet

trim the spaces, delete the columns not related with the analysis

library the required packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(readr)
library(ggplot2)
library(dplyr)
```

loading data

```
superstore <- read_csv("Global_Superstore2.csv")
```

```
## Rows: 51290 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (16): Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Segment, ...
## dbl  (6): Row ID, Sales, Quantity, Discount, Profit, Shipping Cost
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

data process

clean data

```
cleaned_superstore <- superstore %>% clean_names()

head(cleaned_superstore)
```

```
## # A tibble: 6 x 22
##   row_id order_id order_date ship_date ship_mode customer_id segment city  state
##    <dbl> <chr>    <chr>      <chr>     <chr>     <chr>       <chr>   <chr> <chr>
## 1  32298 CA-2012~ 31-07-2012 31-07-20~ Same Day  RH-19495    Consum~ New ~ New ~
## 2  26341 IN-2013~ 05-02-2013 07-02-20~ Second C~ JR-16210    Corpor~ Woll~ New ~
## 3  25330 IN-2013~ 17-10-2013 18-10-20~ First Cl~ CR-12730    Consum~ Bris~ Quee~
## 4  13524 ES-2013~ 28-01-2013 30-01-20~ First Cl~ KM-16375    Home O~ Berl~ Berl~
## 5  47221 SG-2013~ 05-11-2013 06-11-20~ Same Day  RH-9495     Consum~ Dakar Dakar
## 6  22732 IN-2013~ 28-06-2013 01-07-20~ Second C~ JM-15655    Corpor~ Sydn~ New ~
## # ... with 13 more variables: country <chr>, market <chr>, region <chr>,
## #   product_id <chr>, category <chr>, sub_category <chr>, product_name <chr>,
## #   sales <dbl>, quantity <dbl>, discount <dbl>, profit <dbl>,
## #   shipping_cost <dbl>, order_priority <chr>
```

add columns

```
cleaned_superstore$day <- format(as.Date(cleaned_superstore$order_date, "%d-%m-%Y"), "%d")
cleaned_superstore$month <- format(as.Date(cleaned_superstore$order_date,"%d-%m-%Y"), "%m")
cleaned_superstore$year <- format(as.Date(cleaned_superstore$order_date, "%d-%m-%Y"),"%Y")
```

data analysis

frequency of purchase for each customer

```
cleaned_superstore_summary1 <- cleaned_superstore%>%group_by(customer_id, year) %>%
  summarize(number_of_order_id=n()) %>%
  arrange(desc(number_of_order_id))
```

```
## `summarise()` has grouped output by 'customer_id'. You can override using the
## `.groups` argument.
```

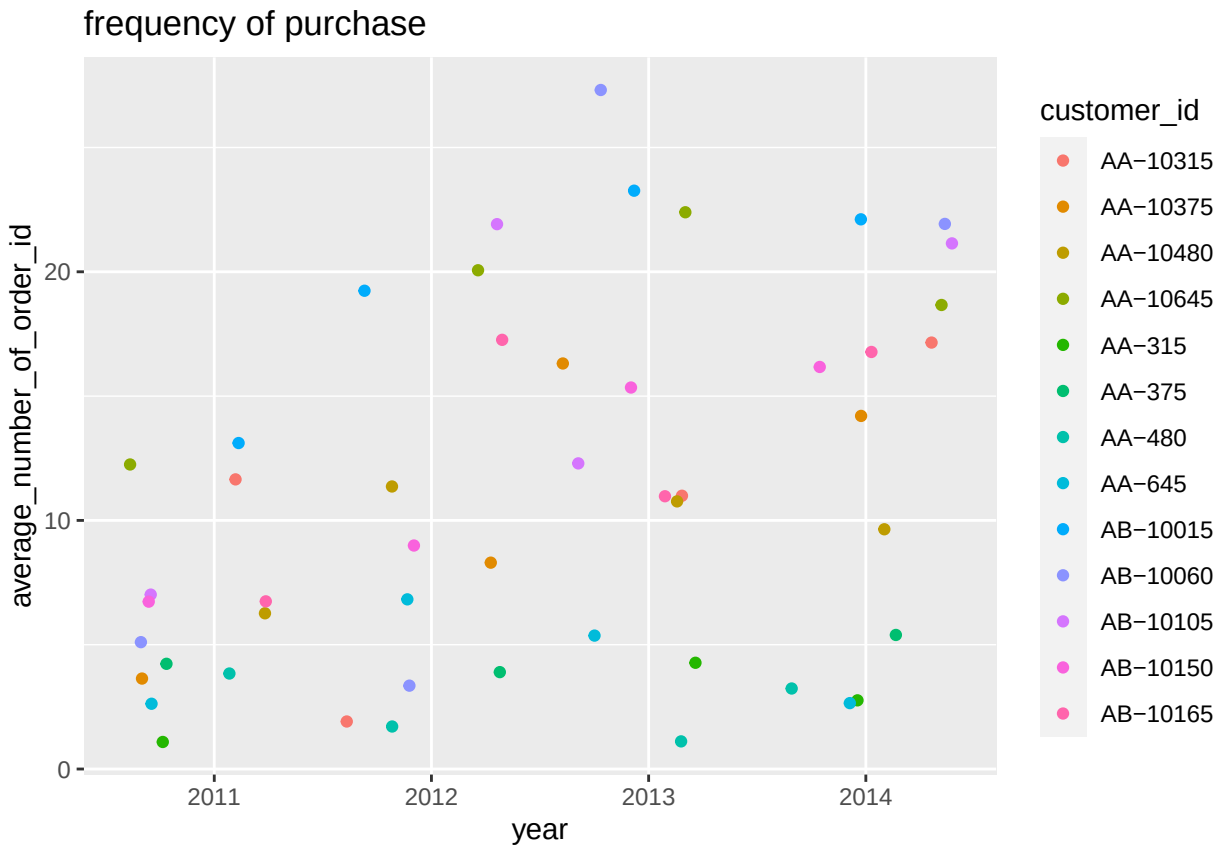
```
ccleaned_superstore_summary1 <- cleaned_superstore_summary1 %>% group_by(customer_id, year) %>%
  summarize(average_number_of_order_id=mean(number_of_order_id))
```

```
## `summarise()` has grouped output by 'customer_id'. You can override using the
## `.groups` argument.
```

```
head_ccleaned_superstore_summary1 <- head(ccleaned_superstore_summary1, n=50)
```

data visualization

```
ggplot(head_ccleaned_superstore_summary1)+
  geom_jitter(mapping=aes(x=year, y=average_number_of_order_id, color=customer_id))+
  labs(title="frequency of purchase ")
```



calculate high frequent(set top 15 as high frequent) customers are contributing more customers

```
cleaned_superstore_summary2 <- cleaned_superstore %>%
  group_by(customer_id, month) %>%
  summarize(number_of_order_id=n(), sum_sales=sum(sales)) %>%
  arrange(desc(sum_sales)) %>%
  arrange(desc(number_of_order_id))
```

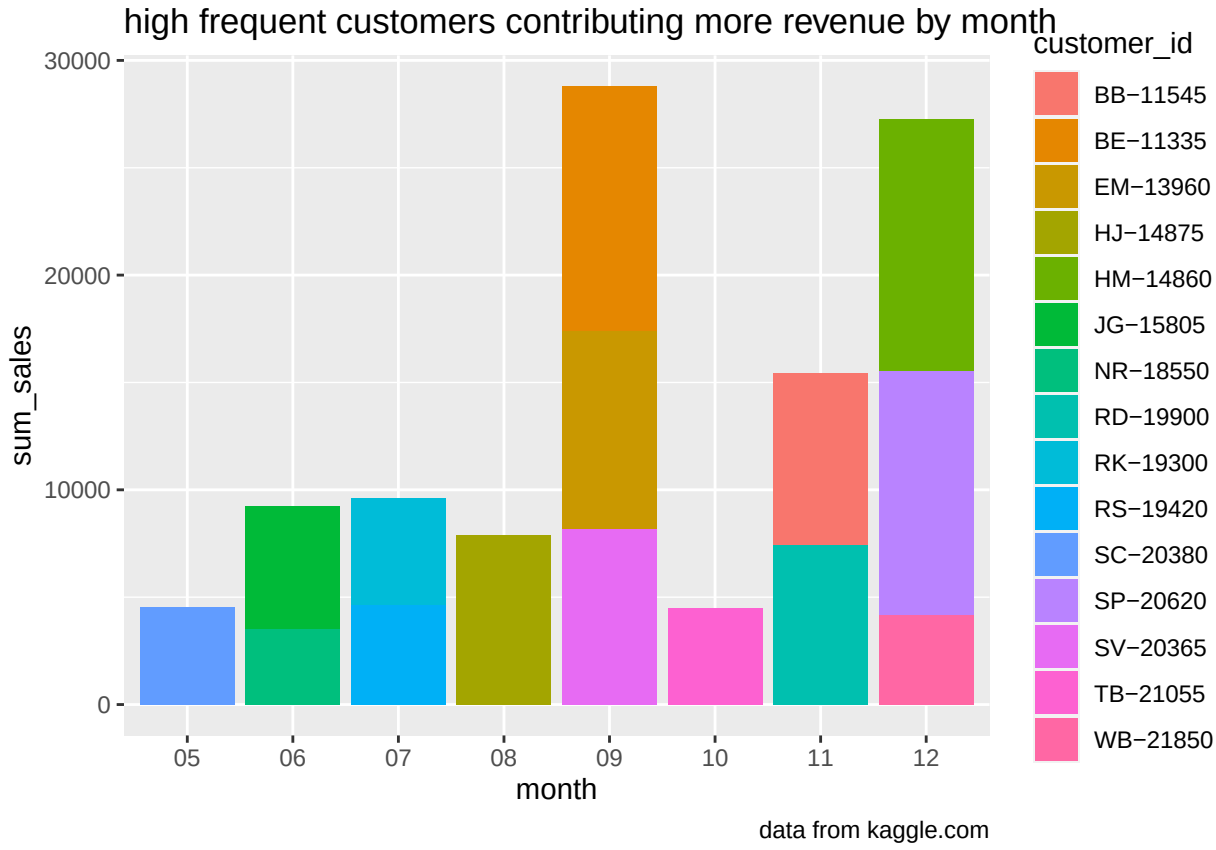
```
## `summarise()` has grouped output by 'customer_id'. You can override using the
## `.groups` argument.
```

set high frequent customers as top 15 of customers

```
head_cleaned_superstore_summary2 <- head(cleaned_superstore_summary2, n=15)
```

data visualization

```
ggplot(data=head_cleaned_superstore_summary2)+
  geom_col(mapping=aes(x=month, y=sum_sales, fill=customer_id))+
  labs(title="high frequent customers contributing more revenue by month", caption="data from kaggle.co
```



Calculate the profit margin group by customer id

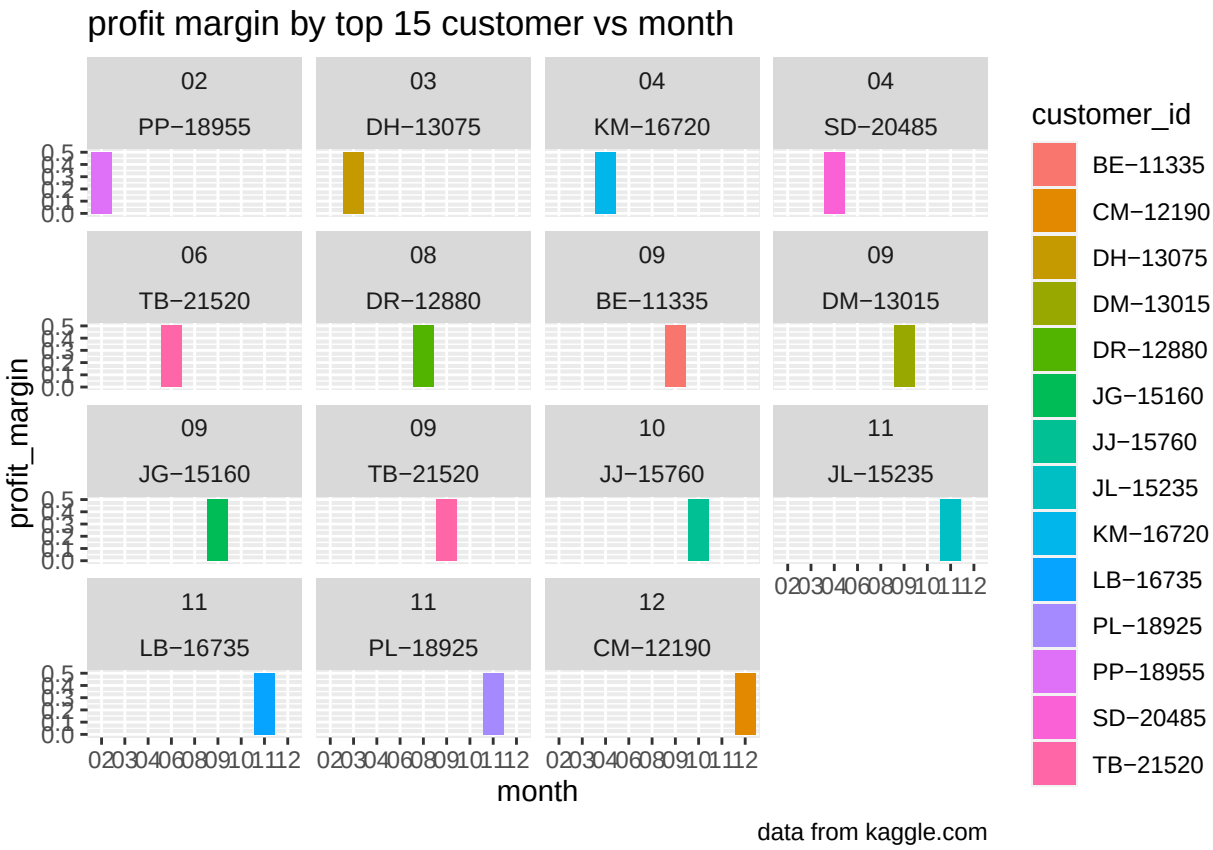
```
cleaned_superstore_summary3 <- cleaned_superstore %>%
    group_by(customer_id, month) %>%
   summarize(number_of_order_id=n(), sum_sales=sum(sales), profit_margin=(profit/sales)) %>%
   arrange(desc(sum_sales)) %>%
   arrange(desc(number_of_order_id)) %>%
   arrange(desc(profit_margin))
```

```
## `summarise()` has grouped output by 'customer_id', 'month'. You can override
## using the `.groups` argument.
```

```
head_cleaned_superstore_summary3 <- head(cleaned_superstore_summary3, n=15)
```

data visualization

```
ggplot(data=head_cleaned_superstore_summary3)+
  geom_col(mapping=aes(x=month, y=profit_margin, fill=customer_id))+
  labs(title="profit margin by top 15 customer vs month", caption="data from kaggle.com")+
  facet_wrap(month~customer_id)
```



calculate which customer segment is most profitable each year

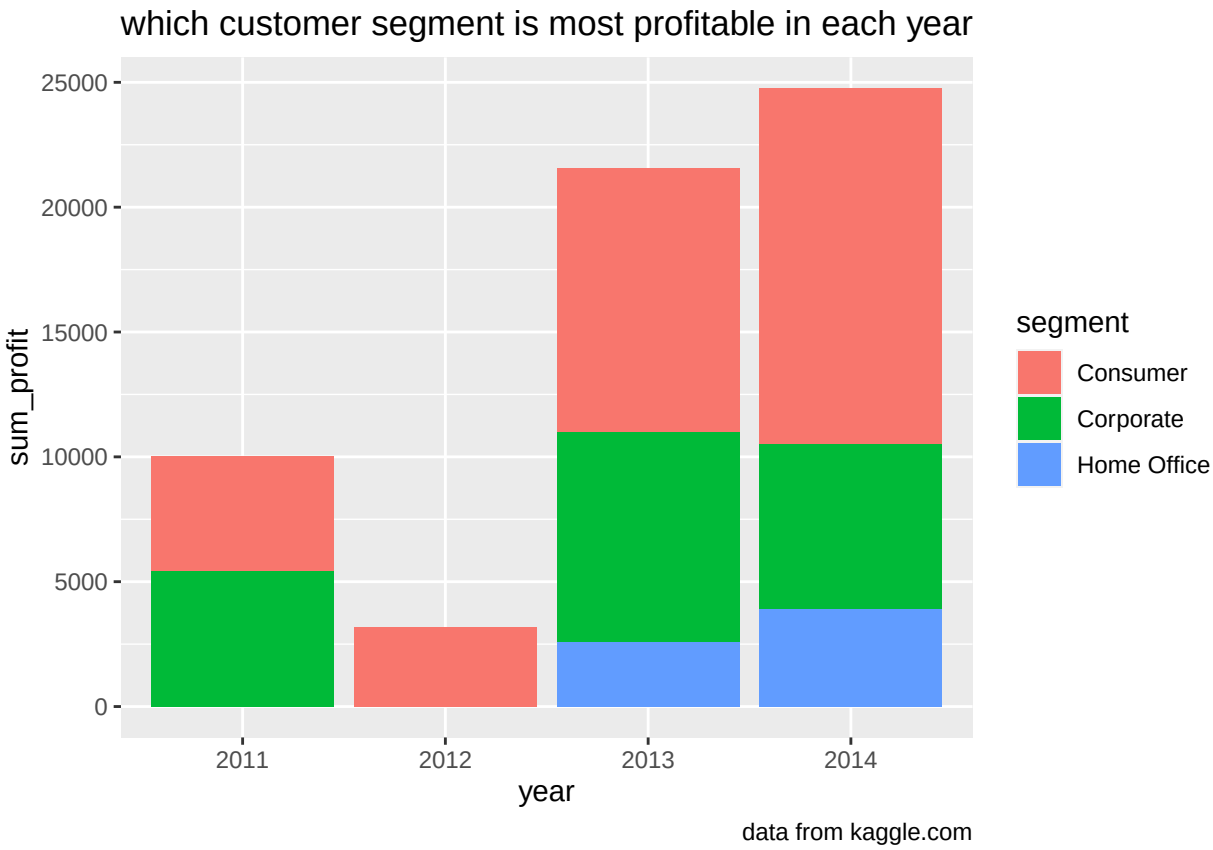
```
cleaned_superstore_summary4 <- cleaned_superstore %>%
  group_by(segment, profit, year) %>%
  summarize(number_of_order_id=n(), sum_profit=sum(profit), profit_margin=(profit/sales))%>%   arrange(
  arrange(desc(sum_profit))
```

```
## `summarise()` has grouped output by 'segment', 'profit', 'year'. You can
## override using the `.groups` argument.
```

```
head_cleaned_superstore_summary4 <- head(cleaned_superstore_summary4, n=15)
```

data visualization

```
ggplot(data=head_cleaned_superstore_summary4)+
  geom_col(mapping=aes(x=year, y=sum_profit, fill=segment))+
  labs(title="which customer segment is most profitable in each year", caption="data from kaggle.com")
```


which customer segment is most profitable in each year
25000
20000
15000
10000
5000
0
sum_profit
2011
2012
2013
2014
year
data from kaggle.com
segment
Consumer
Corporate
Home Office

how customers are distributed accross the countries